FPN++: A SIMPLE BASELINE FOR PEDESTRIAN DETECTION

Junhao Hu[†] *Lei Jin*[†] *Shenghua Gao*[*]

ShanghaiTech University, Shanghai, China
{hujh, jinlei, gaoshh }@shanghaitech.edu.cn

ABSTRACT

Our observation shows that pedestrians' heights greatly affect pedestrian detection performance, and for small pedestrians, their context information is useful for localizing and recognizing these pedestrians. Based on our observation, a FPN++ framework, which is an extension of Feature Pyramid Network (FPN) is proposed. It improves the FPN from the following aspects: i) we modify the backbone of FPN by reducing the stride of convolution from 2 to 1 in FPN from earlier layers, which allows the network to detect smaller pedestrians with more semantically meaningful features extracted from deeper layers, then we replace the convolution with dilated convolution to increase the local receptive fields and facilitate the detection on pedestrians of all scales; ii) a context-aware detection module is introduced in the predictor head of FPN to leverage context information for detection. Extensive experiments on the CityPersons and Caltech pedestrian datasets show that our FPN++ achieves state-of-the-art performance and significantly improves the performance for small pedestrians. Our solution can be readily extended to other detection tasks, and experiments on the VOC2007 benchmark also validate the effectiveness of our solution.

Index Terms— Pedestrian Detection, Feature Pyramid Network

1. INTRODUCTION

Pedestrian detection has been well studied because of its potential applications in autonomous driving, robotics and intelligent surveillance. Recently performance of pedestrian detection has been greatly boosted benefiting from deep learning based approaches [1], but it still suffers the large scale variance of pedestrians [2, 3, 4].

In light of the capability of Feature Pyramid Network (FPN) for detecting objects with different sizes [5], FPN has also been introduced for pedestrian detection. Specifically, in FPN, low-resolution, semantically stronger features are combined with high-resolution, semantically less features with a feature pyramid architecture for objects of different scales. However, as shown in Fig. 1, the performance still drops significantly for small pedestrians. For example, for the pedestrians with height less than 50 pixels, miss rate is around 70%. The poor detection performance of small-scale pedestrians is possibly due to the following reasons: i) the weak representation of small-scale pedestrians in shallow stages [2]; ii) the low context information in Region-of-Interest (RoI) pooling for bounding box representation. As shown in Fig. 1, the central area of ROI features has strong pattern for recognition for large-size instances, while the context helps recognize the pattern of small-scale pedestrians. An intuitive explanation to this phenomenon is that large pedestrians have enough local information on themselves to be recognized, while surrounding context is essential for detection on obscure or dim pedestrians.

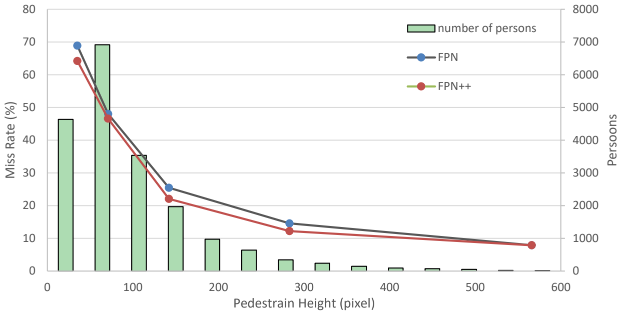


Fig. 1. Green columns demonstrate the height distribution of pedestrians in CityPersons Dataset, ranging from 20 to 600 pixels. Blue lines and red lines show the corresponding miss rate of Feature Pyramid Network (FPN) and our FPN++. We are able to improve the detection performance for small-scale persons, while preserving the accuracy for large-scale persons.

To remedy the deficiency of FPN for small-scale pedestrian detection, we propose to improve the backbone and predictor head from the following aspects: i) we modify the backbone of FPN by reducing the stride of convolution in FPN from earlier layers, which allows the network to detect smaller pedestrians with more semantically meaningful features extracted from deeper layers, and by replacing the convolution with dilated convolution, which increases the lo-

[†] Equal Contribution
[*] Corresponding author

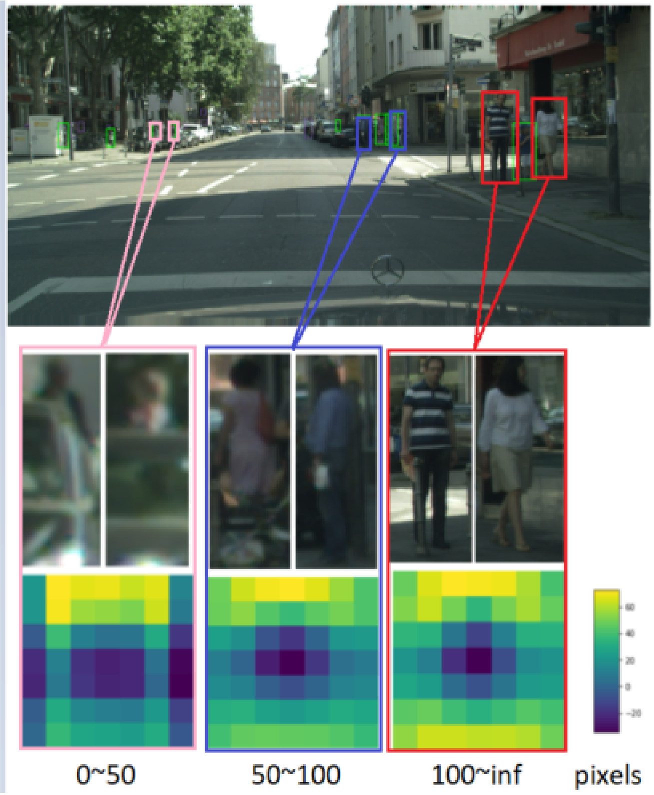


Fig. 2. Analyzed features of pedestrians. Different from the larger ones which have strong activation on central area, small-size pedestrians (left) have strong activation on both sides of the feature map.

cal receptive fields and facilitates the detection on pedestrians of all scales. We denote the FPN with such modification as FPN+; ii) a context-aware detection module is introduced in the predictor head of FPN to leverage context information for classification and bounding box regression. In this way, our solution can well detect both small-scale and large-scale pedestrians. Further, the idea of our FPN++ can be readily extended to the detection of other general objects.

The main contribution of this paper can be summarized as follows:

i) We conduct experiments to analyze the feature distribution of pedestrians with different sizes, and show that context information is vital for the detection of small-size pedestrians.

ii) A FPN++ framework is proposed, which extends the backbone and predictor head respectively.

iii) Extensive experiments on both pedestrian detection and general object recognition validate the effectiveness of our approach.

2. RELATED WORK

Earlier work uses hand-crafted features, including HOG, LBP and LUV [6, 7], for pedestrian detections. Readers may refer to the survey paper[8] to get an overview of hand-crafted features based pedestrian detection. Here we only briefly review deep learning based pedestrian detection methods.

Faster R-CNN [9] is a prevailing method for object detection. Features are extracted on a 16x downsampled feature map. Undoubtedly, it's unsuitable for small-size pedestrians. [10] and [1] both remove the last pooling layer to reduce the original stride to a half. One difference between [10] and [1] is that [1] replaces the common convolution after the removed pooling layers with dilated convolution, while [10] keeps the original design in this part.

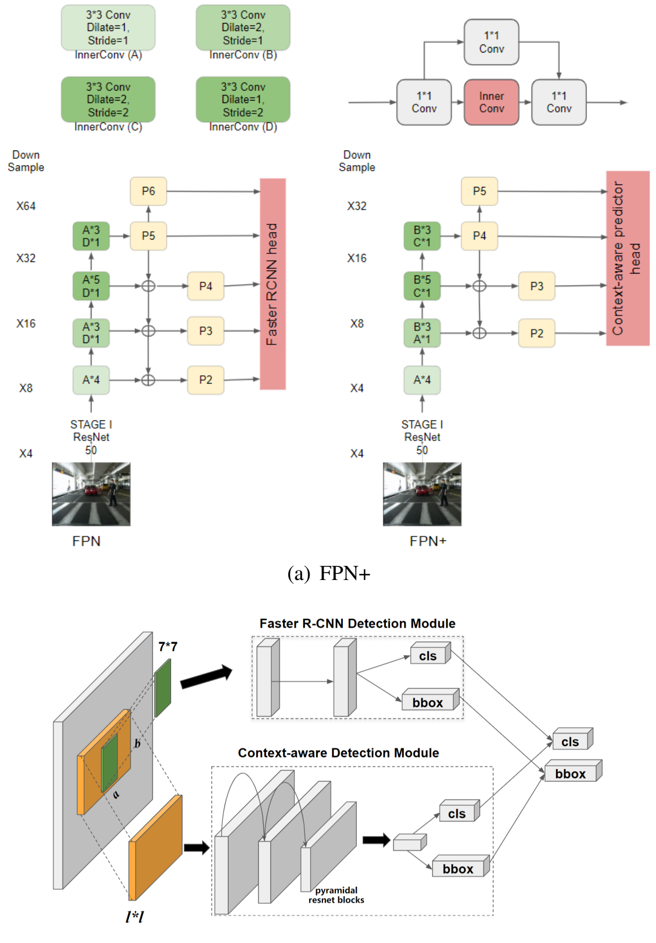
Nevertheless, one feature map cannot satisfy the large scale variance of pedestrians. MSCNN [11] extracts features on feature maps with different spacial resolutions to detect different-scale objects respectively. To improve semantic levels of feature maps from low-level stages, FPN [5] and RON [12] fuse features from neighbouring two stages so that detection sub-network can utilize both higher-resolution feature and higher-semantic feature. However, [2] argue that the semantic level of small pedestrians is still not high enough. They propose a MHN solution which processes features from low-level stages into high-level semantic features by deep multi-branch networks.

3. APPROACH

Our FPN++ customizes the FPN for more robust pedestrian detection. It consists of two parts: i) a FPN+ backbone, which we reduce the stride of convolution and replace it with dilated convolution; and ii) a predictor head which leverages a context-aware predictor head that utilizes context information around the RoI for pedestrian detection and bounding box regression, consequently improves the small-scale pedestrian detection. Our overall network structure is demonstrated in Fig. 3.

3.1. FPN+ backbone

The bottom right part of Fig 3(a) demonstrates a basic structure of our network. Our FPN+ extends FPN by reducing the stride from an early stage to enlarge resolution of our feature map, and replacing the convolution with dilated convolutions to enlarge local receptive field of high level features. Specifically, we change the stride from 2 to 1 in the first block of the stage II, consequently feature maps from stage II, III and IV in our FPN+ are 2^2 times as large as the original ones in FPN. With the expansion of spatial resolution, we detect the smallest pedestrians in stage III and combine the context from stage III and stage IV so that our detection sub-network gets features with higher semantic level. Similarly, the improve-



(a) FPN+

(b) Context-aware predictor head

Fig. 3. Overall architecture of our network. 3(a) left is the original FPN and right is the newly proposed backbone FPN++ for better semantics. 3(b) is a context-aware predictor head for environmental information. Given an input image, it first goes through FPN+ and extracts features on four stages. After a naive foreground and background classification and localization, it's then fed into the context-aware predictor head for a second time classification and localization.

ment is also beneficial to detecting pedestrians of medium and large scales detected in deeper stages. As the improvement of semantic level and extra higher-level context are from the top-to-down structure, there is little need to use the multi-branch structure as what MHD [2] does, consequently our FPN+ reduces large amount of computation and memory occupation.

After the changing of the spacial resolution, we dilate [13] all of the convolution filters from stage III to stage V by 2 (except the convolution in the first block of stage III) to keep the receptive fields of every pixel as large as the ones in the original FPN. This allows the feature maps in low stages to be used both for lower-stage detection and higher-stage detection, which is another internal condition to avoid multi-branch structure. Although the receptive fields of every pixel is expanded by dilation, there are large overlap in two adjacent pixels' receptive fields. So the undilated RPN's receptive fields in stage i from the improved backbone are closer to the ones in stage $i-1$ from the original backbone, rather than stage i. It meets our design intent to detect smaller pedestrians in higher stages.

3.2. Context-aware predictor head

To utilize surrounding contextual information, we design a sub-network (see Fig. 3(b)) with meticulous care. First, each RoI is expanded by α times on width and β times on height. Then, this larger RoI is pooled into a $l \times l$ feature map. Two modified pyramidal resnet blocks are applied to further enhance the feature performance. Each block consists of a 1×1 conv to reduce dimensionality, a 3×3 conv dilated by 3, and another 1×1 conv to recover dimensionality. All of the convolutional layers are of stride 1 and no padding. Since padding is not used, the size of input is larger than the one of output in each block. We design these no-padding pyramidal block to avoid noise caused by padding because our post-RoI-pooling feature is much smaller than usual feature map. In each block, we crop the input feature for the block and get the center area so that we can add it to the output of the block. Finally, we apply an RPN (a 3×3 conv followed by two sibling 1×1 convs) on the output to perform object/nonobject binary classification and bounding box regression. In practice, α is set to 3, β is set to 2 and l is 15. Coupled with dilated convolution, the two ResNet blocks can further enhance the semantic information with surrounding context. Obviously, the post-RoI-Pooling convolution requires much less extra computation and memory resources than convolution on feature maps in backbone.

The prediction of the module described above is combined with the prediction of the original Faster R-CNN. We simply average the two predictions, and the performance is already satisfactory. More complicated combination strategy is possible but it is beyond the study scope of this paper.

4. EXPERIMENTS

4.1. Evaluation metric

We use the log miss rate averaged over the false positive per image (FPPI) range of $[10^{-}2; 10^{0}]$ (MR^2), the official evaluation metric of the CityPersons dataset [10], to measure detection performance. It is the average value of miss rates for 9 FPPI (false positives per image) rates evenly spaced in the log-space ranging from 10^2 to 10^0 (lower score indicates better performance).

4.2. Implementation Details

Our method is implemented on the PyTorch platform. We use a ResNet50 [14] pretrained on the ImageNet[15] dataset as backbone network. All the newly added parameters are initialized with gaussian distribution with standard deviation of 0.01. We optimize the detector with Stochastic Gradient Descent (SGD) with 0.9 momentum and 0.0001 weight decay.

In order to further improve detection performance, we use a denser class of anchor scales, which can be calculated as:

$$s_i = \lambda * s_{\min} * \sqrt{r} * (\frac{s_{\max}}{s_{\min}})^{\frac{2*i-1}{N}}, i = 1, 2, ...N/2$$

where λ is the upsampling factor and r is the ratio of pedestrians. Specifically, $N = 24$ and $r = 0.41$. learning rate is set to 10^{-3} for the first 40k iterations, and reduced by 10^{-1} for the next 20k iterations. Following the convention in existing research [1][10], for the Caltech pedestrian dataset, λ is set to 2, while for CityPersons, λ is set to 1.3.

4.3. The CityPersons dataset

The CityPersons dataset is built upon Cityscapes dataset [16] with a finer annotation of pedestrians. It is the most popular dataset on pedestrians detection in recent years due to its diversity and challenge. It consists of 5000 images with a resolution of 1024x2048 pixels. For sake of fairness and comparisons, we use the original training set and validation set, which are composed of 2,975 and 500 images respectively. The rest are used in testing.

We compare our method with other state-of-the-art pedestrian detection methods, including Repulsion[17], OR-CNN [18], MHN-D[2] and Adapted Faster R-CNN [10]. During training and testing, all images are upsampled to 1.3x. Results of different methods are demonstrated in Table. 1. We can see that our method achieves state-of-the-art performance on the *All* test set without bells and whistles. Specifically, we outperform the best method by 1.81% on *All* subset. That is, our method can handle pedestrians of all heights. Further, we implement three of these networks with both VGG-16 and ResNet50 in Pytorch. For Adapted Faster R-CNN, it needs 14722MB GPU memory with ResNet50. For OR-CNN, it can be inferred from its structure that it needs more memory than Adated Faster R-CNN because it adds complicated detection structure on Adated Faster R-CNN. While our network requires less memory resources than others with the ResNet50 backbone because we make full use of the semantic fusion structure of FPN.

Method	backbone	upsample	All	Memory (MB)
Adapted Faster RCNN	VGG16	1.3x	43.86	3689
MHN-D	VGG16	1.3x	39.16	19908
OR-CNN	VGG16	1.3x	40.19	–
Repulsion Loss	ResNet50	1.5x	39.17	22895
Ours	ResNet50	1.3x	**37.36**	**14543**

Table 1. Results on the CityPersons *test* set. Lower is better. We are able to reach state-of-the-art performance on *All* test set. This demonstrates the effectiveness of our method to handle pedestrians of different heights and different levels of occlusions. Compared to other state-of-the-art methods, we are able to improve by 1.81% on *All*.

4.4. The Caltech pedestrian dataset

The Caltech pedestrian dataset [19] consists of a 2.5-hour video recorded on a bus on the streets of Los Angeles. It's divided into a total of 11 videos. We follow the standard criteria [19], where the training samples are extracted every 3 frames on the first 6 videos, and testing are extracted every 30 frames on the other 5 videos.

Following the standards from [2], we use reasonable all test set (pedestrians over 20 pixels). Note that we upsample the images twice during both training and testing, which is a common trick to further improve our performance. Results and ROC are summarized in Fig. 4. We achieve the lowest miss rate, which validates the effectiveness of our method.

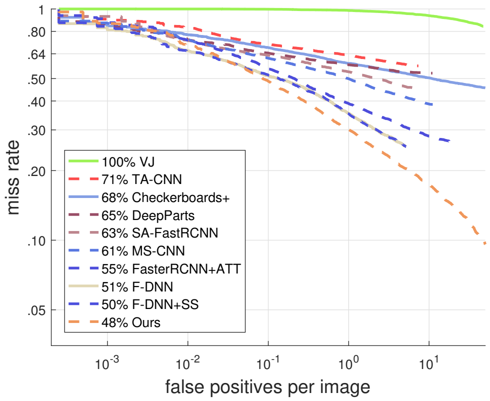


Fig. 4. Results and ROC on Caltech pedestrian detection. Pedestrians over 20 pixels are used for evaluation, which is a harder criteria. We achieve the lowest miss rate.

4.5. Ablation study

4.5.1. Evaluation of FPN+ module

Two related and representative feature pyramid methods are compared to our FPN+ module on the Citypersons *valiadation* dataset to demonstrate the effectiveness of FPN++. One of them is FPN [5], the other is a modified FPN where P3 to P6 in FPN are used to detect corresponding anchors originally detected on P2 to P5 respectively. Because of the similarity with backbone in RetinaNet [20], we denote the second comparison as modified FPN (mFPN).

The result in Table. 2 shows our FPN+ achieves better performance on all the scales than baselines. With same spacial resolution on each scale, FPN++ shows advantage over

FPN, which demonstrates the importance of higher-level semantics. With same-level semantics on each scale, FPN+ has lower MR than mFPN, which proves the importance of larger resolution.

Height	30	50	100	200	400	ALL
FPN	68.90	48.03	25.49	14.58	7.90	42.32
mFPN	75.06	52.41	25.10	13.60	9.00	42.89
FPN+	**64.23**	**46.57**	**22.07**	**12.21**	**7.87**	**39.36**

Table 2. Detection results evaluated on the CityPersons *validation* set. All modules are trained on training set. Our proposed method achieves the best performance.

4.5.2. *Evaluation of FPN+ on VOC2007*

Our FPN+ is a general module which can be readily adapted for other general object detection task. We propose to evaluate our FPN+ with the VOC2007 benchmark [21] for general object detection. VOC2007 contains 20 categories and a total of 9963 images, 5011 for training and 4952 for testing. Mean average precision (mAP) is used as the evaluation criterion.

The comparison between our FPN+ and FPN [5] are shown in Table 3. We can see that our FPN+ outpeforms FPN which validates the effectiveness of our method. Note that FPN is not method corresponding to the best performance on VOC2007. We just want to demonstrate that, with a similar configuration and parameters, our networks is the more effective for object detection.

4.5.3. *Comparison of three block structure*

To figure out whether dilation is beneficial to performance and what size of receptive field is better when spacial resolution is enlarged, we compare three backbones with different structures of the blocks as following:

(1)**FPN+ with no dilation (ND)**: use the original block in ResNet50 with common convolution filters.

(2) **FPN+ with half dilation (HD)**: dilates half of the 3×3 convolution filters from stage III to stage V by 2 and keeps the other half unchanged.

(3)**FPN+ with all dilation (AD)**: dilates all the 3×3 convolution filters from stage III to stage V by 2, which is our FPN+.

To avoid the potential impact from pre-trained weights on these experiments, we train these three backbones from scratch in this experiment. The results are shown in Table. 4. It can be seen that detection performance goes better in all of the scales with the ratio of dilated filters increasing. Including from the results, dilated convolution is beneficial to detection on small, medium and large scales. We conjecture that this is because receptive field exercises as a great influence on semantics no matter what size of detected objects is.

Method	Backbone	Stage	Performance (mAP)
FPN	ResNet50	P2, 3, 4, 5, 6	73.1
FPN+	ResNet50	P3, 4, 5, 6, 7	**74.5**

Table 3. Results on VOC2007 *test* set.

Height	30	50	100	200	400	ALL
ND	82.00	64.86	44.48	36.84	23.01	61.46
HD	76.26	58.86	43.92	35.42	21.69	60.18
AD	73.66	56.06	36.37	28.02	18.24	**53.59**

Table 4. FPN+ results on different blocks, which are trained from scratch.

Height	30	50	100	200	400	ALL	sub-network parameters
FPN+	65.47	46.01	22.98	**12.45**	**7.46**	38.85	13.25M
FPN++	**62.82**	**42.45**	**22.11**	13.24	8.32	**37.82**	13.26M
FPN++ w. 2FC	65.80	44.97	**23.11**	14.34	9.28	39.56	26.51M

Table 5. Detection results and detection sub-network parameter sizes corresponding to different configurations of detection sub-network.

4.5.4. *Evaluation of context-aware predictor head*

To utilize surrounding context for pedestrian detection, we design a context-aware predictor head. We show the results of FPN+ where vanilla detection module is used and our FPN++ in Table 5. Compared with FPN+, FPN++ achieves better performance which mainly comes from the promotion on small-size pedestrian detection (HS 2.6 points lower and H50 3.5 points lower). Usually, small objects are more likely to be blurred or dim, so this comparison can prove that our surrounding context detection works as we expects since the dataset does not contain annotations on fuzzy degree.

4.5.5. *Comparison of different structures of context-aware predictor head.*

[22] completely duplicate the structure of the original Faster R-CNN module into a parallel module. However, in our opinion, the two FC layers in the original sub-networks are difficult to train and bring a large number of parameters. To verify our opinion, we design a 2FC sub-network. We pool the larger RoI into a 7 × 7 feature map and send it to a FC sub-network exactly same as the other one. The predictions from two parallel sub-network are merged at last. We denote this baseline as FPN++ w. 2FC. In Table 5, FPN++ w. 2FC shows no advantage over FPN++, even over FPN+. The degrade can be attributed to the large amount of extra parameters. By contrast, our CEM bring an obvious promotion at the cost of little extra parameters.

5. CONCLUSION

We have studied the impact of resolution, semantic level and receptive field on FPN based pedestrian detection and presented a simple but effective backbone for pedestrian detection. Also, we have explored the difference between features from different sizes of pedestrians and designed a context-aware predictor head. Extensive experiments validate the effectiveness of our solution for pedestrian detection.

6. REFERENCES

[1] Liliang Zhang, Liang Lin, Xiaodan Liang, and Kaiming He, "Is faster r-cnn doing well for pedestrian detection?," in *ECCV*. Springer, 2016, pp. 443–457.

[2] Jiale Cao, Yanwei Pang, and Xuelong Li, "Exploring multi-branch and high-level semantic networks for improving pedestrian detection," *arXiv preprint arXiv:1804.00872*, 2018.

[3] Shanshan Zhang, Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele, "How far are we from solving pedestrian detection?," in *CVPR*, 2016, pp. 1259–1267.

[4] Jiayuan Mao, Tete Xiao, Yuning Jiang, and Zhimin Cao, "What can help pedestrian detection?," in *CVPR*. IEEE, 2017, pp. 6034–6043.

[5] Tsung-Yi Lin, Piotr Dollár, Ross B Girshick, Kaiming He, Bharath Hariharan, and Serge J Belongie, "Feature pyramid networks for object detection.," in *CVPR*, 2017, vol. 1, p. 4.

[6] Xiaoyu Wang, Tony X Han, and Shuicheng Yan, "An hog-lbp human detector with partial occlusion handling," in *ICCV*. IEEE, 2009, pp. 32–39.

[7] Woonhyun Nam, Piotr Dollár, and Joon Hee Han, "Local decorrelation for improved pedestrian detection," in *Advances in Neural Information Processing Systems*, 2014, pp. 424–432.

[8] Markus Enzweiler and Dariu M Gavrila, "Monocular pedestrian detection: Survey and experiments," *TPAMI*, , no. 12, pp. 2179–2195, 2008.

[9] Ross Girshick, "Fast r-cnn," in *ICCV*, 2015, pp. 1440–1448.

[10] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele, "Citypersons: A diverse dataset for pedestrian detection," in *CVPR*, 2017, vol. 1, p. 3.

[11] Zhaowei Cai, Quanfu Fan, Rogerio S Feris, and Nuno Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *ECCV*. Springer, 2016, pp. 354–370.

[12] Tao Kong, Fuchun Sun, Anbang Yao, Huaping Liu, Ming Lu, and Yurong Chen, "Ron: Reverse connection with objectness prior networks for object detection," in *CVPR*, 2017, vol. 1, p. 2.

[13] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *TPAMI*, vol. 40, no. 4, pp. 834–848, 2018.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., "Imagenet large scale visual recognition challenge," *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.

[16] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele, "The cityscapes dataset for semantic urban scene understanding," in *CVPR*, 2016, pp. 3213–3223.

[17] Xinlong Wang, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, and Chunhua Shen, "Repulsion loss: Detecting pedestrians in a crowd," in *CVPR*, June 2018.

[18] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z Li, "Occlusion-aware r-cnn: Detecting pedestrians in a crowd," in *ECCV*. Springer, 2018, pp. 657–674.

[19] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona, "Pedestrian detection: An evaluation of the state of the art," *PAMI*, vol. 34, 2012.

[20] Tsung-Yi Lin, Priyal Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," *TPAMI*, 2018.

[21] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman, "The pascal visual object classes (voc) challenge," *IJCV*, vol. 88, no. 2, pp. 303–338, 2010.

[22] Jianan Li, Xiaodan Liang, ShengMei Shen, Tingfa Xu, Jiashi Feng, and Shuicheng Yan, "Scale-aware fast r-cnn for pedestrian detection," *IEEE Transactions on Multimedia*, vol. 20, no. 4, pp. 985–996, 2018.